A modeling approach for designing explainable Artificial Intelligence

Álvaro Navarro[1,*], Ana Lavalle[1], Alejandro Maté[1] and Juan Trujillo[1]

[1]*Lucentia Research Group, Department of Software and Computing Systems, University of Alicante, Carretera San Vicente del Raspeig s/n, San Vicente del Raspeig, 03690, Alicante, Spain*

Abstract

EXplainable Artificial Intelligence (XAI) has become one of the most important and complex issues to address in the current Artificial Intelligence (AI) field. This topic is the key point to foster the adoption of the AI solutions, providing reliability in the AI systems. There are multiple XAI techniques, with its own characteristics such as the technical level or confidence required to make use of them. In this context, modeling approaches can aid in managing the complexity and describe the implementation alternatives available. However, there is a lack of these modeling approaches that help the XAI developers to select and apply the most adequate XAI techniques in the specific scenarios. Motivated by this issue, we aim to present a modeling approach to support XAI developers in the XAI processes. Then, we propose a conceptual model, which is represented in a Unified Modeling Language (UML) class diagram, to capture the key XAI elements. The main advantage of our proposed conceptual model is that it takes into consideration: (i) how the explanations should be generated; (ii) how these explanations will be combined and supported; and (iii) what explainable interfaces, where these explanations are included, will be generated and presented to the end-users. Finally, to test the applicability of our proposal, we have exemplified it through a real project focused on diagnosing the Attention-Deficit/Hyperactivity Disorder (ADHD).

Keywords

EXplainable Artificial Intelligence, Conceptual Modeling, Attention-Deficit/Hyperactivity Disorder

1. Introduction

Currently, our society heavily depends on Artificial Intelligence (AI) systems [1]. Moreover, the industry 4.0 aims to increase it, improving the citizens' everyday life through these systems, which encompass Machine Learning (ML) and Deep Learning (DL) models. These systems and models learn and execute different tasks in different areas such as medical diagnosis [2] or flood forecasting [3]. Nonetheless, many of these systems are opaque, *i.e.*, it is difficult to understand the reasons behind the decisions that have been made by them.

This opacity prevents end-users from making use of their right to have explainable systems

ER2023: Companion Proceedings of the 42nd International Conference on Conceptual Modeling: ER Forum, 7th SCME, Project Exhibitions, Posters and Demos, and Doctoral Consortium, November 06-09, 2023, Lisbon, Portugal

[*] Corresponding author.

alvaro.nl@dlsi.ua.es (Á. Navarro); alavalle@dlsi.ua.es (A. Lavalle); amate@dlsi.ua.es (A. Maté); jtrujillo@dlsi.ua.es (J. Trujillo)

0000-0002-9730-2106 (Á. Navarro); 0000-0002-8399-4666 (A. Lavalle); 0000-0001-7770-3693 (A. Maté); 0000-0003-0139-6724 (J. Trujillo)

according to the European AI perspective [4], where it is argued that the AI systems must be understandable and explain how have reached their decisions. Moreover, this black-box nature can also lead these systems to make unfair decisions before being detected. For example, there have been AI systems that have developed a discriminatory behaviour toward some races (In [5], an AI system was used by judges to decide if a person is kept in prison or not) or genders (In [6] an AI system from Amazon to qualify job applicants was skewed against women). Both above-presented points emphasize the relevance of understanding the AI systems decision-making processes before reaching society.

Given this demand for interpretable and understandable systems, eXplainable Artificial Intelligence (XAI) emerged to address the lack of information in AI systems, as a key aspect in the integration of AI in different domains. In this context, XAI aims to provide additional information for the opaque models in their current state. On the one hand, it allows ML developers (henceforth referred to as ML experts) to develop ML and DL models, which are subsets of the AI systems, in a better way. This is possible by verifying that these models are free from inconsistencies, errors, and skewed behaviours. On the other hand, it allows end-users to (i) understand the output of the models, (ii) trust in the rationale or rules learned by the models, and (iii) have confidence on the decisions made by these models.

However, applying explainability techniques is a difficult task to be achieved. First, the most correct XAI techniques vary depending on different dimensions involved in the case study, which creates a complex scenario that needs to be deeply analyzed. Second, there is a lack of conceptual models to help XAI developers/designers (henceforth referred to as XAI experts) to be capable of applying XAI techniques to different AI systems to generate the necessary explanations.

In such a complex scenario, although explainable approaches are being generated in different AI scenarios, they are not providing end-users with the information to build trust through explainable interfaces [7].

Consequently, the XAI processes heavily depend on XAI experts to (i) analyse the scenario, (ii) interpret the key elements in the specific case study, and (iii) translate the applied XAI techniques in an explainable interface.

In the context of XAI, these problems could be solved. Specifically, it is possible to solve them by creating a conceptual model that allows to specify all the XAI points and analyse what is the most appropriate explainable techniques and representation in each case study.

Therefore, it is essential to provide XAI experts with tools that, together with their XAI expertise, allow them to design and analyze XAI solutions to be implemented, so that users obtain the right explanations for the specific AI contexts where XAI will be incorporated.

Nevertheless, to define this conceptual model is far from trivial. In order to achieve it and guide the XAI processes, there are different dimensions that should be included. Moreover, several XAI State-Of-The-Art (SOTA) works (*e.g.*, [8]) have emphasized that explanations can support explanations, which permits to help end-users by generating more complete and understandable explainable interfaces [7]. Apart from this, in the XAI SOTA, there are being discussed many points that are not yet so clear, which presents a novel and partly-unexplored field that should be mapped by proposals that help to understand the contexts of application of XAI.

Motivated by this, we aim to provide a modeling approach that achieve these points. Thus, we propose a conceptual model represented in a Unified Modeling Language (UML) [9] class diagram that aids XAI experts in designing XAI systems, which captures the key elements in the XAI field. Thanks to this, it is possible to draw the steps to achieve the most adequate explanations adapting them to the specific case studies, which will be presented below.

First, our proposed conceptual model captures the different kinds of explanations to be generated. Second, it captures how these explanations will be combined and supported to present a complete explainable interface to end-users. Finally, this conceptual model takes into consideration how to design the explainable interface, where the different explanations will be included [7].

The rest of the paper is structured as follows. Section 2 presents the main concepts of XAI and UML that are included in this paper. Section 3 presents the related works in the area. Section 4 presents the proposed conceptual model for XAI. The case study where our proposal has been applied is presented in Section 5. Finally, the conclusions and the future work are presented in Section 6.

2. Background

As previously-argued, many of the current AI systems present a lack of transparency and interpretability, the relevance of XAI in the current AI field is clear. In this way, we will summarize the key XAI concepts to the adequate understanding of our proposal. Moreover, since our proposal is represented in an UML class diagram, the UML key concepts will be also presented. Hence, both XAI and UML main concepts will be presented, as follows.

First, there is presented a clear distinction among different models in the XAI field. This dichotomy is widely accepted as a classification between transparent models, which present an enough degree of interpretability, and opaque models (also categorized such as model-agnostic or model-agnostic), where we should apply XAI techniques to understand their decision-making processes [8].

Second, an XAI technique is categorized depending on its scope, its methodology, and its usage [10]. In the context of the usage categories, there are intrinsic techniques, which aim to decode specific models, and post-hoc techniques, which aim to understand the logic behind agnostic models. Specifically, there are different post-hoc techniques defined [8]: (i) simplification techniques, which aim to create a more simple and understandable model that presents a similar performance than the original; (ii) text techniques, which tackle the explainability by giving information information abut the learning model process; (iii) visual techniques, which aim at observing the model's behavior through visualizations; (iv) local techniques, which segment the solution space and give explanations that are relevant for the whole model; (v) feature relevance techniques, which aim to extract the most relevant features that the model have taken into consideration to make its decisions and weights these features; and (vi) example explanation techniques, which are based on the extraction of data examples which refer to the output generated by the model that should be explained. Moreover, it is possible to define custom explanations by building new techniques and combine different techniques to present a more complete explanation.

Third, there is emphasized the relevance of identifying the different end-users in XAI. Thus, it is possible to adapt the explainable interface, where the selected explanations will be included, for them in the specific case study [7],

Finally, as our proposal is represented in an UML class diagram, we will describe the main concepts of UML that we have included in this paper. On the one hand, an UML *class* provides information to create objects and contains their attributes, an UML *abstract class* plays a role of superclass but can not be instantiated, and an UML *enumeration* that represents the possible value of an attribute in a fixed set of discrete values. On the other hand, the UML *association* denotes the relationship between different UML classes in a unidirectional or bidirectional way, the UML *composition* is a stronger relationship where an object of one class contains another class (if the container object is destroyed, the contained objects are also destroyed), and the UML *generalization* that represents the inheritance between classes where a subclass inherits attributes and operations from a superclass.

Once presented these XAI and UML concepts, we are able to present the related work in the XAI area, our proposal, and the case study where it has been applied, as follows.

3. Related work

In this section, we will discus (i) how XAI is being applied in the AI field, (ii) the different possibilities to model XAI, and (iii) how our proposal supports XAI in its current state by bridging the gap between the XAI and modeling fields, as follows.

As argued in [8], the XAI applications, which are being applied in different domains (*e.g.*, healthcare [11]), directly depend on the DL or ML models which the explanations are based on. In this context, there are different frameworks focused on applying XAI in Natural Language Processing [12], Reinforcement Learning [13], or even in analysing the existing goals, metrics and users to reveal the necessity of making an effort to drive the design and implementation of XAI [7].

Analyzing the modeling field, we have taken into consideration different techniques. On the one hand, the i-star framework [14] has been applied in different contexts such as ML [15]. Hence, it could help to extract the XAI requirements in AI scenarios. On the other hand, different Object Constraint Language (OCL) rules [16] have been defined in different case studies (*e.g.*, Data Warehouses [17] or Visualizations [18]). However, as this paper aims to cover all XAI key elements in a Platform Independent Model, the above-presented techniques are out of this paper scope. Finally, a recent work studied the impact of applying ML and UML in the system analysis and design context [19].

Once presented these works, we can observe that current XAI practice lacks adequate tools and techniques to allow XAI experts to design their solutions and ensuring that the most adequate techniques are selected while exploring potential alternatives. Moreover, this does not depend on the specific context or dimensions where is applied, *i.e.*, to define a conceptual model is crucial to achieve the goal that each field faces.

However, none of these works have addressed the modeling of XAI to correctly include it in AI scenarios, which is essential to help XAI experts to select the most adequate explainable techniques. Motivated by this unreached goal, we present a modeling approach to help the XAI experts to select the most adequate explanations in the specific AI contexts, which will be presented the next section.

4. A modeling approach to define explanations

In this section, we present our proposed conceptual model, which captures the key elements included when XAI is applied in AI scenarios. The aim of this metamodel, which is aligned with the UML [9] technique standardization, is allowing XAI experts to model and analyze the different explanations to be generated, and AI models involved, in order to design and select the most adequate XAI techniques and interfaces for each specific case study. In order to present this conceptual model (Fig. 1[1]), which helps the XAI experts to decide which of them are more adequate, we have meticulously studied the SOTA of XAI (*e.g.*, [8, 10]). In order to completely present our proposed conceptual model, its different points of our proposal will be presented below.

First, our conceptual model captures the different kinds of explanations. This point takes into consideration the different dimensions involved to correctly select the most appropriate techniques depending on the specific case study.

Second, this proposed model also captures how these explanations will be combined and supported to present a complete explainable interface to the end user. Hence, the resulting explanations generated will be more useful to end-users.

Finally, it takes into consideration how to design the explainable interface, where the different explanations will be included. Thus, end-users will be provided with a more complex a complete interface to observe and interact, which is essential to help them to understand the reasons behind the decisions made by the AI system. In this context, the different XAI techniques selected will be included in the explainable interface [7].

As a result of the introduction of the proposed XAI conceptual model, we obtain an XAI design analysis process that helps the XAI expert to correctly select the most adequate explainable techniques for the specific case studies.

Therefore, our proposal lets the XAI experts analyze the case study where XAI will be applied. Consequently, we present a conceptual model that includes the relevant elements in XAI scenarios, which is presented in Figure 1. It shows how to design the explanations given different elements.

In the following, the proposed elements included in our conceptual model will be explained. As we can observe in Fig. 1, where this conceptual model is presented, there are three different type of elements included: classes (presented in a yellow color), abstract classes (presented in a blue color), and enumerations (presented in a green color). Moreover, we have divided the proposed elements in three different blocks to help to understand and apply our proposal. Specifically, there are the explanation, technique and model blocks, which will be presented below.

First, the explanation block is presented. This block includes the “Explanation” abstract class that captures the simple or multimodal (also called combined) explanation generated. Moreover, this block also captures the “DomainKnowledge” class that supports the explanation through different ways (included in the “DomainKnowledgeResourceEnum” enumeration: ontologies, dictionaries and domain experts). The format of the explanations is also relevant in the “TextFormat” and “VisualFormat” classes that indicate how the explanation will be presented. Moreover,

[1] Image in different formats located in repository

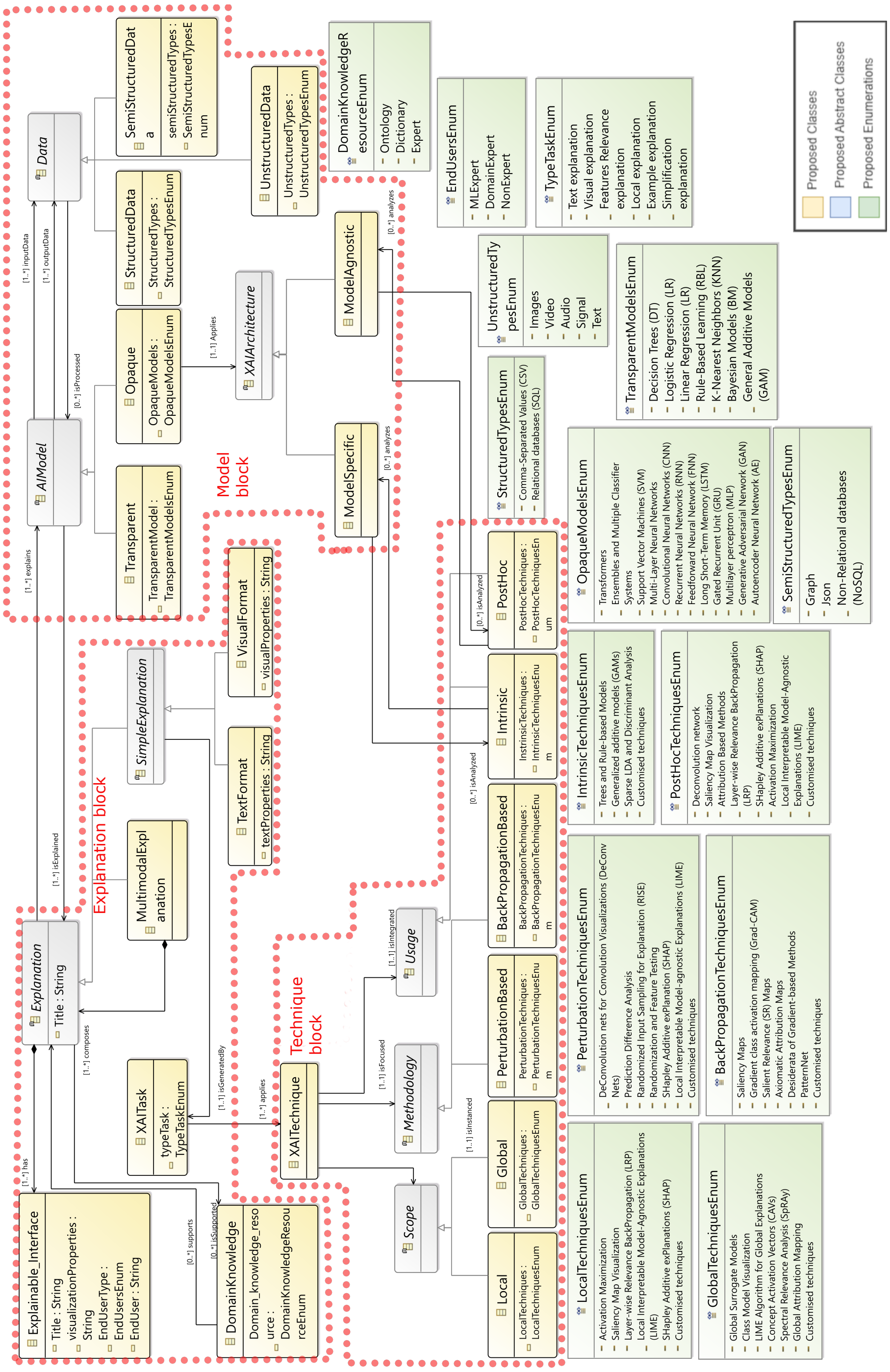

Data
AIModel
Explanation
Title : String
Explainable_Interface
Title : String
visualizationProperties : String
EndUserType : EndUsersEnum
EndUser : String
DomainKnowledge
Domain_knowledge_resource : DomainKnowledgeResourceEnum
XAITask
typeTask : TypeTaskEnum
XAITechnique
MultimodalExplanation
SimpleExplanation
VisualFormat
visualProperties : String
TextFormat
textProperties : String
Transparent
TransparentModel : TransparentModelsEnum
Opaque
OpaqueModels : OpaqueModelsEnum
StructuredData
StructuredTypes : StructuredTypesEnum
SemiStructuredData
semiStructuredTypes : SemiStructuredTypesEnum
UnstructuredData
UnstructuredTypes : UnstructuredTypesEnum
XAIArchitecture
ModelAgnostic
ModelSpecific
Usage
Methodology
Scope
PostHoc
PostHocTechniques : PostHocTechniquesEnum
Intrinsic
InstrinsicTechniques : IntrinsicTechniquesEnum
BackPropagationBased
BackPropagationTechniques : BackPropagationTechniquesEnum
PerturbationBased
PerturbationTechniques : PerturbationTechniquesEnum
Global
GlobalTechniques : GlobalTechniquesEnum
Local
LocalTechniques : LocalTechniquesEnum
[1..*] inputData
[1..*] outputData
[0..*] isProcessed
[1..*] explains
[1..*] isExplained
[1..*] composes
[1..*] has
[0..*] supports
[0..*] isSupported
[1..1] isGeneratedBy
[1..*] applies
[1..1] Applies
[0..*] analyzes
[0..*] isAnalyzed
[1..1] isIntegrated
[1..1] isFocused
[1..1] isInstanced
Model block
Explanation block
Technique block
DomainKnowledgeResourceEnum
- Ontology
- Dictionary
- Expert
EndUsersEnum
- MLExpert
- DomainExpert
- NonExpert
TypeTaskEnum
- Text explanation
- Visual explanation
- Features Relevance explanation
- Local explanation
- Example explanation
- Simplification explanation
UnstructuredTypesEnum
- Images
- Video
- Audio
- Signal
- Text
StructuredTypesEnum
- Comma-Separated Values (CSV)
- Relational databases (SQL)
TransparentModelsEnum
- Decision Trees (DT)
- Logistic Regression (LR)
- Linear Regression (LR)
- Rule-Based Learning (RBL)
- K-Nearest Neighbors (KNN)
- Bayesian Models (BM)
- General Additive Models (GAM)
OpaqueModelsEnum
- Transformers
- Ensembles and Multiple Classifier Systems
- Support Vector Machines (SVM)
- Multi-Layer Neural Networks
- Convolutional Neural Networks (CNN)
- Recurrent Neural Network (RNN)
- Feedforward Neural Network (FNN)
- Long Short-Term Memory (LSTM)
- Gated Recurrent Unit (GRU)
- Multilayer perceptron (MLP)
- Generative Adversarial Nerwork (GAN)
- Autoencoder Neural Network (AE)
SemiStructuredTypesEnum
- Graph
- Json
- Non-Relational databases (NoSQL)
IntrinsicTechniquesEnum
- Trees and Rule-based Models
- Generalized additive models (GAMs)
- Sparse LDA and Discriminant Analysis
- Customised techniques
PostHocTechniquesEnum
- Deconvolution network
- Saliency Map Visualization
- Attribution Based Methods
- Layer-wise Relevance BackPropagation (LRP)
- SHapley Additive exPlanations (SHAP)
- Activation Maximization
- Local Interpretable Model-Agnostic Explanations (LIME)
- Customised techniques
PerturbationTechniquesEnum
- DeConvolution nets for Convolution Visualizations (DeConv Nets)
- Prediction Difference Analysis
- Randomized Input Sampling for Explanation (RISE)
- Randomization and Feature Testing
- SHapley Additive exPlanation (SHAP)
- Local Interpretable Model-agnostic Explanations (LIME)
- Customised techniques
BackPropagationTechniquesEnum
- Saliency Maps
- Gradient class activation mapping (Grad-CAM)
- Salient Relevance (SR) Maps
- Axiomatic Attribution Maps
- Desiderata of Gradient-based Methods
- PatternNet
- Customised techniques
LocalTechniquesEnum
- Activation Maximization
- Saliency Map Visualization
- Layer-wise Relevance BackPropagation (LRP)
- Local Interpretable Model-Agnostic Explanations (LIME)
- SHapley Additive exPlanations (SHAP)
- Customised techniques
GlobalTechniquesEnum
- Global Surrogate Models
- Class Model Visualization
- LIME Algorithm for Global Explanations
- Concept Activation Vectors (CAVs)
- Spectral Relevance Analysis (SpRAy)
- Global Attribution Mapping
- Customised techniques
Proposed Classes
Proposed Abstract Classes
Proposed Enumerations

Figure 1: Proposed conceptual model for capturing the key XAI elements.

there is also included the “XAITask” class. In this class, we have defined different types of tasks that can be represented. In this context, we propose to isolate the different previously-presented post-hoc techniques (section 2) to provide this new concept of XAI tasks, which is closely linked to the XAI techniques concept (the definition of these techniques is included in the proposed conceptual model and will be presented in the next block). Specifically, this new concept generates the explanations thanks to the application of the most appropriate XAI techniques. Hence, the possible values for a type of task, which are presented in the "TypeTaskEnum" enumeration, are [8]: text, visual, feature relevance, local, example, and simplification. Finally, this block presents the "ExplainableInterface" class that contains the different explanations generated by one XAI task. This class that takes into consideration the specific end-user (which can be an ML expert, domain expert, or non-expert) and the visualization properties of the final format of the explainable interface. Detailing the cardinalities, an explainable interface is build by at least one explanation and can contain a lot of explanations; the domain knowledge can support zero or many explanations (and, transitively, the explainable interface) and these explanations can be supported by zero or many domain knowledge resources.

Second, the technique block is focused on categorizing the different XAI techniques, where an above-presented XAI task applies one or more of these XAI techniques. In this context, an XAI technique always has a scope, which differentiates between techniques focused on the whole model or a subspace of this (global or local, respectively); a methodology, which refers to how an explanation works (perturbation [20] or backpropagation [21, 22] based methods); and an usage, which categorizes the techniques between the focused on specific AI models' architectures or without depending on them (intrinsic or post-hoc, respectively). Due to space constraints, we will not detail all techniques included in these categories, which are presented in the proposed conceptual model (Fig. 1). Moreover, we should specify that the different above-presented categories can be combined depending on the specific XAI technique. For example, the Local Interpretable Model-Agnostic Explanations (LIME) technique [23] is perturbation based, post-hoc and could be local or global depending of the specific application.

Third, the model block contains information about the AI model that should be explained. In this context, the "AIModel" abstract class is provided with one or more type of the input data and, after processing the data and executing different steps, provides one or more types of the output data. Both input and output data are represented in the abstract class "Data" and can be instanced as: structured data (.csv files, etc.), semi-structured data (.json files, etc.) and unstructured data (images, etc.). Moreover, the AI model can be transparent or opaque [8]. On the one hand, the transparent models are: Decision Trees (DT), Logistic Regression (LR), Linear Regression (LR), K-Nearest Neighbors (KNN), Bayesian Models (BM), and General Additive Models (GAM). On the other hand, there are different opaque models (Convolutional Neural Networks (CNN), Transformers, etc.) that refer to AI architectures that can be categorized as model-specific, which are analyzed by intrinsic techniques, and model-agnostic, which are analyzed by post-hoc techniques.

Finally, once modeled the previously-describe elements and enumerations, we achieve a conceptual model that helps the XAI experts to apply XAI in AI scenarios. Hence, our UML-standardized proposal lets them to cover the different dimensions included in the XAI processes, which is essential to select the most adequate XAI approaches in the specific case studies. In the next section our case of application will be presented

5. Case study

In order to test the applicability of our proposal, we have applied it in a real case study. Specifically, it has been applied to an existing project on AI-driven Attention-Deficit/Hyperactivity Disorder (ADHD) diagnosis and treatment project, called Balladeer[2]. Therefore, in this section we will present (i) the different steps and elements of our proposal specified to this scenario, and (ii) the final explainable interface achieved thanks to having applied our proposal, as follows.

5.1. Conceptual modeling to define the explainable interface in the ADHD context

In the following, the different captured elements in the ADHD case study, which allow to define the explainable interface and its explanations, will be presented in Fig. 2[3].

First, the AI model presents an opaque nature that aims to diagnose the patients as positive or negative ADHD cases. Specifically, it is a CNN model-agnostic architecture that processes unstructured 2D Electroencephalography (EEG) signals -captured by using the emotiv headset [24]- from the patients as input data, and provides the results in a structured data format (.csv files). Moreover, this model consist of six parallel blocks of two one-dimensional (1D) kernels that are applied on cascading, where each kernel presents a different length to extract features at different frequencies and executes the elu activation function. Moreover, each kernel also presents a number of 64 filters, a same padding mode and a number of 15 epochs. In order to diagnose the ADHD for each patient, there is applied an average pool between these six blocks and, after executing it, the sigmoid function is also applied to return a value of 0 (ADHD negative case) or 1 (ADHD positive case). This above-described model architecture is shown in Fig. 3.

Second, the XAI expert aims to decode the black-box nature of the CNN model. In this context, the XAI expert -the first author of this paper plays this role-, who has made use of the previous-presented metamodel (Fig. 1), defines three different explanations that should be included in the explainable interface. More specifically, these three explanations are defined such as local, perturbation based and post-hoc [10], taking into consideration its scope, methodology and usage, respectively. Moreover, two of them compose a multimodal explanation represented in visual formats, and the other one is presented in a text format. These explanations are presented below.

The first explanation is a feature relevance XAI task. In order to achieve it, the SHAP [25] technique has been selected as the most adequate, which is inside the sub-space of the intersection of the local, perturbation and post-hoc XAI techniques spaces. Hence, we can extract the most relevant brain sensors, *i.e.*, which brain zones have presented a higher activity during the tests applied and their events.

The second explanation is a visual XAI task. In order to apply it, a customised technique is implemented. Hence, there is presented a headplot where the more relevant sensors, provided by the previous explanation Thus, there is applied a segmentation where the more relevant sensors are emphasized in a red colour.

[2] https://balladeer.lucentia.es/en/home-2/

[3] Image in different formats located in repository

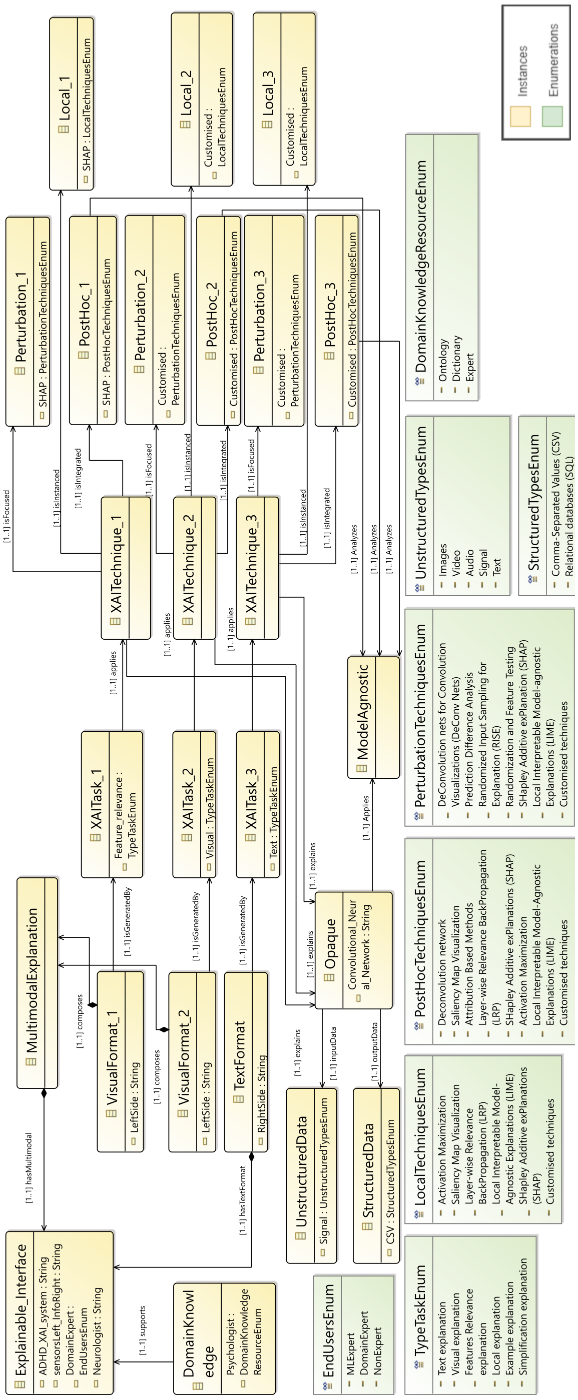

Instances
Enumerations
Explainable_Interface
ADHD_XAI_system : String
sensorsLeft_InfoRight : String
DomainExpert : EndUsersEnum
Neurologist : String
DomainKnowledge
Psychologist : DomainKnowledgeResourceEnum
[1..1] supports
[1..1] hasMultimodal
[1..1] hasTextFormat
MultimodalExplanation
VisualFormat_1
LeftSide : String
VisualFormat_2
LeftSide : String
TextFormat
RightSide : String
[1..1] composes
XAITask_1
Feature_relevance : TypeTaskEnum
XAITask_2
Visual : TypeTaskEnum
XAITask_3
Text : TypeTaskEnum
[1..1] isGeneratedBy
XAITechnique_1
XAITechnique_2
XAITechnique_3
[1..1] applies
[1..1] isFocused
[1..1] isInstanced
[1..1] isIntegrated
Perturbation_1
SHAP : PerturbationTechniquesEnum
PostHoc_1
SHAP : PostHocTechniquesEnum
Perturbation_2
Customised : PerturbationTechniquesEnum
PostHoc_2
Customised : PostHocTechniquesEnum
Perturbation_3
Customised : PerturbationTechniquesEnum
PostHoc_3
Customised : PostHocTechniquesEnum
Local_1
SHAP : LocalTechniquesEnum
Local_2
Customised : LocalTechniquesEnum
Local_3
Customised : LocalTechniquesEnum
ModelAgnostic
[1..1] Analyzes
[1..1] Applies
Opaque
Convolutional_Neural_Network : String
[1..1] explains
[1..1] inputData
[1..1] outputData
UnstructuredData
Signal : UnstructuredTypesEnum
StructuredData
CSV : StructuredTypesEnum
EndUsersEnum
MLExpert
DomainExpert
NonExpert
TypeTaskEnum
Text explanation
Visual explanation
Features Relevance explanation
Local explanation
Example explanation
Simplification explanation
LocalTechniquesEnum
Activation Maximization
Saliency Map Visualization
Layer-wise Relevance BackPropagation (LRP)
Local Interpretable Model-Agnostic Explanations (LIME)
SHapley Additive exPlanations (SHAP)
Customised techniques
PostHocTechniquesEnum
Deconvolution network
Saliency Map Visualization
Attribution Based Methods
Layer-wise Relevance BackPropagation (LRP)
SHapley Additive exPlanations (SHAP)
Activation Maximization
Local Interpretable Model-Agnostic Explanations (LIME)
Customised techniques
PerturbationTechniquesEnum
DeConvolution nets for Convolution Visualizations (DeConv Nets)
Prediction Difference Analysis
Randomized Input Sampling for Explanation (RISE)
Randomization and Feature Testing
SHapley Additive exPlanation (SHAP)
Local Interpretable Model-agnostic Explanations (LIME)
Customised techniques
UnstructuredTypesEnum
Images
Video
Audio
Signal
Text
StructuredTypesEnum
Comma-Separated Values (CSV)
Relational databases (SQL)
DomainKnowledgeResourceEnum
Ontology
Dictionary
Expert

Figure 2: Application of the proposed conceptual model in the ADHD case study.

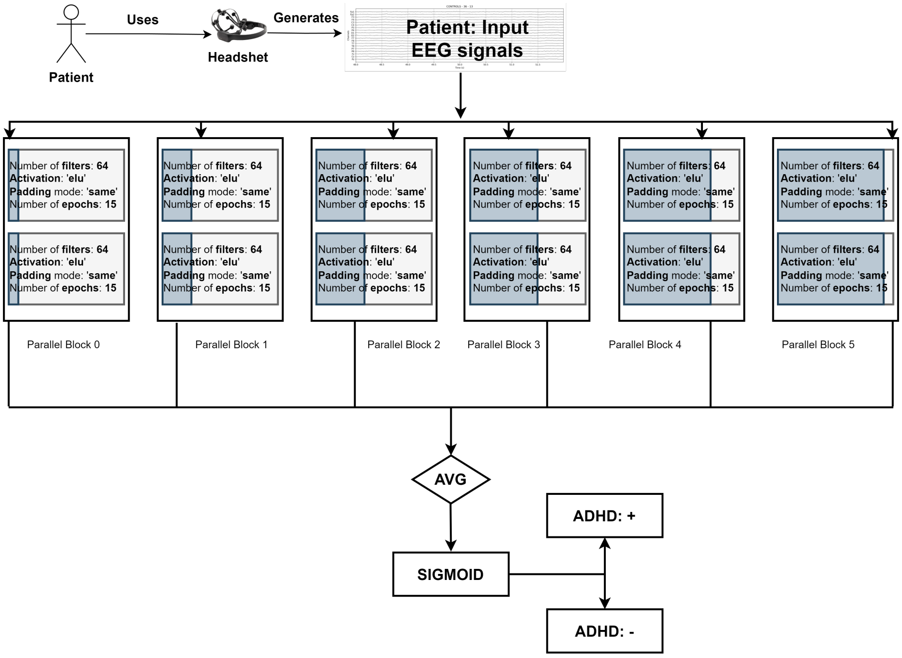


Figure 3: AI model architecture.

The third explanation is a text XAI task. This explanations translates the patients information to an organized text distribution, which lets end-users to take this information into consideration through generating key words, sentences a paragraphs.

Finally, all above-presented points will be included in the explainable interface. This explainable interface aims to help the neurologists, who are the specific end-users as domain experts, to understand the logic behind the AI model decision-making process. Moreover, this explainable interface will be organized in two parts: the left side for the headplot, where the sensor relevance is included, and the right side for the patient information. Consequently, this explainable interface will help the neurologists involved in the case study. In the following, this explainable interface will be presented.

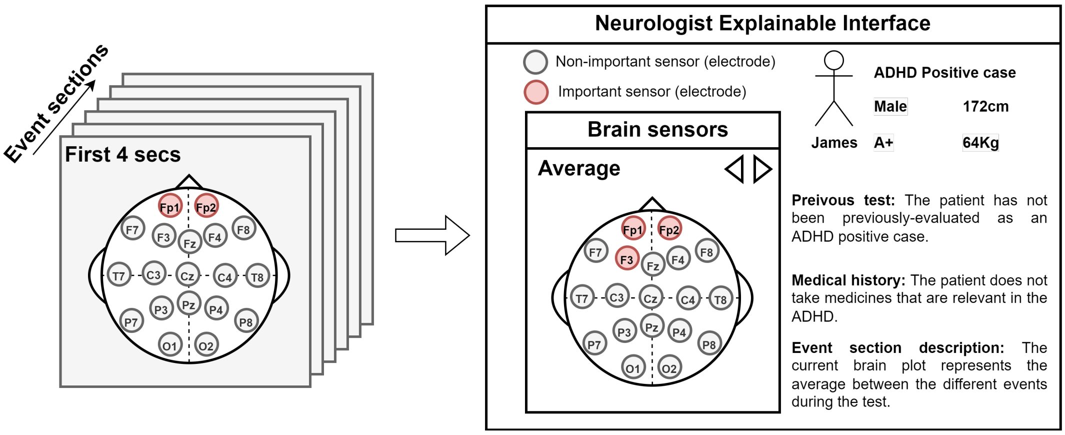
5.2. Explainable interface

Following the above-presented points, the explanations can be implemented. We then show the final explainable interface achieved, which aims to support the end-users: the neurologists.

As presented in Fig. 4, the different headplots extracted from the different layers -before the feature vectors are executed-, are included in the explainable interface. In this interface, it is possible to navigate to observe the different headplots. Moreover, the patient's information captured apart from the AI system (weight, blood group, etc.) is also included.

In this context, the sensors relevance presented in the headplots play a feature relevance role -each sensor is each feature-, which is supported by a visual explanation -different colours in the headplots-. Furthermore, there are different headplots to analyze depending on the specific time point or event for the patient. On the other hand, there is patient's information that can help the neurologists to understand the diagnose and verify it: patient's previous diagnoses, blood group, etc.

In Fig. 4, we can observe the explainable interface that supports the neurologists, which has been achieved thanks to our proposal.



Preivous test: The patient has not been previously-evaluated as an ADHD positive case.

Medical history: The patient does not take medicines that are relevant in the ADHD.

Event section description: The current brain plot represents the average between the different events during the test.

Figure 4: Neurologists' explainable interface designed thanks to our proposal.

6. Conclusions and future work

Nowadays, AI systems have an enormous relevance in society, and this relevance is growing day by day. Moreover, these systems are being applied in different fields. In this context, it is necessary to analyse them properly before they reach society.

To make the most of AI systems, it is crucial to provide explanations that can shed light on the reasoning behind these systems decision-making processes, which generate the outputs. This is especially important in advanced AI systems like DL models, which can produce precise results but lack the ability to comprehend the rationale underlying their decisions.

To address this issue, the field of XAI emerged, which offers a range of techniques and options to generate explanations that can supplement the limited transparency of existing AI systems. However, despite the diversity of XAI techniques available, each tailored to explain specific AI models and generate different types of information, there is still a lack of modeling approaches to help the XAI experts to select the most adequate XAI explanations in the specific context, which is essential to achieve comprehensive and effective explanations for end-users.

Aiming to face this problem, in this paper we have bridged the XAI and conceptual modeling fields. To achieve it, we have meticulously studied the most relevant works (*e.g.*, [8, 7, 10]) in the XAI context to extract the key XAI elements that have been captured by our proposal.

Thus, we have presented a conceptual model represented in an UML class diagram, which captures these key elements in the XAI field, which will support the XAI experts to design and implement explanations in AI scenarios. Specifically, our proposed conceptual model takes into consideration (i) how the explanations should be generated; (ii) how these explanations will be combined and supported to present a complete explainable interface to the end-users; and (iii) what explainable interfaces will be generated so that end-users can see and interpret the decision-making processes of the AI systems.

In order to show the applicability of our proposal, we have applied it to an existing project on AI-driven Attention-Deficit/Hyperactivity Disorder (ADHD) diagnosis and treatment, which is a safe-critical area. Given the sensitivity of the information involved and the diverse set of the end-users and recipients in the medical domain, our case study serves as an ideal example where the modeling of XAI should be approached with care to ensure that each end-user receives the appropriate information.

As a result of the application of our proposal, we have been able to design an explainable interface that provides different explanations adapted to the end-users: the neurologists. Thus, it presents (i) the events EEG signals -captured by using the emotiv headset [24]- from the patients, and (ii) the patient's information that can help to understand the ADHD diagnosis result. Consequently, our proposal not only helps the XAI expert to design and implement the explainable interface, but also helps end-users to understand the logic behind the decisions made by the specific AI system.

In future works, our plans are: (i) to explore alternatives to apply the defined approaches (*e.g.*, situational method engineering [26]) (ii) to explore different abstraction levels such as how to take into consideration certain aspects of the inputs for their adequate representation in the explainable interface; due to the possible combinatorics possible between input data, architectures and output data; (iii) to study how to formally define and use properly formed and universal OCL [16] rules, based on the relevant SOTA of XAI, to ensure the correct design of the model and avoid arbitrary linking the proposed elements; and (iv) to improve or make more systematic the derivation of the different abstraction and technical levels to the final interfaces, which will be presented to end-users. In this sense, we intend to carry out an experiment with end-users to validate the different dimensions of the proposal aside from the purely functional aspects. Consequently, (v) we will provide a multi-dimensional model-driven approach [27] that facilitates faster and less costly XAI implementations in a semi-automatic way and also facilitates the maintainability of XAI codes and their subsequent derivation [28, 29]. Finally, (vi) we will provide the guidelines for the proposed models and (vii) apply a more thorough evaluation by testing the final approach with different people, who will interact the XAI interface.

Acknowledgments

This work has been co-funded by (i) the AETHER-UA project (PID2020-112540 RB-C43) funded by the Spanish Ministry of Science and Innovation, and (ii) the BALLADEER (PROMETEO/ 2021/ 088) project funded by the Conselleria de Innovación, Universidades, Ciencia y Sociedad Digital (Generalitat Valenciana).

References

[1] G. Vilone, L. Longo, Explainable artificial intelligence: a systematic review, arXiv preprint arXiv:2006.00093 (2020). doi:https://doi.org/10.48550/arXiv.2006.00093.

[2] N. Amoroso, D. Pomarico, A. Fanizzi, V. Didonna, F. Giotta, D. La Forgia, A. Latorre, A. Monaco, E. Pantaleo, N. Petruzzellis, et al., A roadmap towards breast cancer therapies supported by explainable artificial intelligence, Applied Sciences 11 (2021) 4881. doi:https://doi.org/10.3390/app11114881.

[3] S. Prasanth Kadiyala, W. L. Woo, Flood prediction and analysis on the relevance of features using explainable artificial intelligence, in: 2021 2nd Artificial Intelligence and Complex Systems Conference, 2021, pp. 1–6.

[4] B. Goodman, S. Flaxman, European union regulations on algorithmic decision-making and a “right to explanation”, AI magazine 38 (2017) 50–57. doi:https://doi.org/10.1609/aimag.v38i3.2741.

[5] B. C. C. Office, B. C. S. Office, F. D. of Corrections, ProPublica, Compas recidivism risk score data and analysis, https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis, 2021. Accessed: 2023/08/21.

[6] J. Weissmann, Amazon created a hiring tool using a.i. it immediately started discriminating against women, https://slate.com/business/2018/10/amazon-artificial-intelligence-hiring-discrimination-women.html, 2018. Accessed: 2023/08/21.

[7] S. Mohseni, N. Zarei, E. D. Ragan, A multidisciplinary survey and framework for design and evaluation of explainable ai systems, ACM Transactions on Interactive Intelligent Systems (TiiS) 11 (2021) 1–45. doi:https://doi.org/10.1145/3387166.

[8] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al., Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai, Information fusion 58 (2020) 82–115. doi:https://doi.org/10.1016/j.inffus.2019.12.012.

[9] I. Jacobson, The unified software development process, Pearson Education India, 1999.

[10] A. Das, P. Rad, Opportunities and challenges in explainable artificial intelligence (xai): A survey, arXiv preprint arXiv:2006.11371 (2020). doi:https://doi.org/10.48550/arXiv.2006.11371.

[11] S. Bharati, M. R. H. Mondal, P. Podder, A review on explainable artificial intelligence for healthcare: Why, how, and when?, IEEE Transactions on Artificial Intelligence (2023) 1–15. doi:10.1109/TAI.2023.3266418.

[12] J. Yu, A. I. Cristea, A. Harit, Z. Sun, O. T. Aduragba, L. Shi, N. A. Moubayed, Interaction: A generative xai framework for natural language inference explanations, in: 2022 International Joint Conference on Neural Networks (IJCNN), 2022, pp. 1–8. doi:10.1109/IJCNN55064.2022.9892336.

[13] R. Dazeley, P. Vamplew, F. Cruz, Explainable reinforcement learning for broad-xai: A conceptual framework and survey, 2021. URL: https://arxiv.org/abs/2108.09003. doi:10.48550/ARXIV.2108.09003.

[14] F. Dalpiaz, X. Franch, J. Horkoff, istar 2.0 language guide, arXiv preprint arXiv:1605.07767 (2016).

[15] J. M. Barrera, A. Reina Reina, A. Maté, J. Trujillo, et al., Applying i* in conceptual modelling in machine learning (2021). URL: http://hdl.handle.net/10045/118806.

[16] O. M. G. (OMG), Unified modeling language specification 1.5, http://www.omg.org/cgi-bin/doc?formal/03-03-01, 2003. Accessed: 2023/08/21.

[17] A. Maté, J. Trujillo, Tracing conceptual models' evolution in data warehouses by using the model driven architecture, Computer Standards and Interfaces 36 (2014) 831–843. URL: https://www.sciencedirect.com/science/article/pii/S0920548914000075. doi:https://doi.org/10.1016/j.csi.2014.01.004.

[18] A. Lavalle, A. Maté, J. Trujillo, Requirements-driven visualizations for big data analytics: A model-driven approach, in: A. H. F. Laender, B. Pernici, E.-P. Lim, J. P. M. de Oliveira (Eds.), Conceptual Modeling, Springer International Publishing, Cham, 2019, pp. 78–92. doi:https://doi.org/10.1007/978-3-030-33223-5_8.

[19] A. Gadhi, R. M. Gondu, C. M. Bandaru, K. C. Reddy, O. Abiona, Applying uml and machine learning to enhance system analysis and design, International Journal of Communications, Network and System Sciences 16 (2023) 67–76. doi:10.4236/ijcns.2023.165005.

[20] M. Ivanovs, R. Kadikis, K. Ozols, Perturbation-based methods for explaining deep neural networks: A survey, Pattern Recognition Letters 150 (2021) 228–234. URL: https://www.sciencedirect.com/science/article/pii/S0167865521002440. doi:https://doi.org/10.1016/j.patrec.2021.06.030.

[21] R. Rojas, The Backpropagation Algorithm, Springer Berlin Heidelberg, Berlin, Heidelberg, 1996, pp. 149–182. URL: https://doi.org/10.1007/978-3-642-61068-4_7. doi:10.1007/978-3-642-61068-4_7.

[22] A. Bhat, A. S. Assoa, A. Raychowdhury, Gradient backpropagation based feature attribution to enable explainable-ai on the edge, in: 2022 IFIP/IEEE 30th International Conference on Very Large Scale Integration (VLSI-SoC), 2022, pp. 1–6. doi:10.1109/VLSI-SoC54400.2022.9939601.

[23] M. T. Ribeiro, S. Singh, C. Guestrin, "why should i trust you?": Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, Association for Computing Machinery, New York, NY, USA, 2016, p. 1135–1144. URL: https://doi.org/10.1145/2939672.2939778. doi:10.1145/2939672.2939778.

[24] M. Duvinage, T. Castermans, M. Petieau, T. Hoellinger, G. Cheron, T. Dutoit, Performance of the emotiv epoc headset for p300-based applications, Biomedical engineering online 12 (2013) 1–15. doi:https://doi.org/10.1186/1475-925X-12-56.

[25] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions 30 (2017). URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

[26] B. Henderson-Sellers, J. Ralyté, Situational method engineering: state-of-the-art review, Journal of Universal Computer Science (2010). doi:http://hdl.handle.net/10453/13456.

[27] O. M. G. (OMG), Model driven architecture guide rev. 2.0, https://www.omg.org/cgi-bin/doc?ormsc/14-06-01, 2014. Accessed: 2023/08/21.

[28] M. Edwards, S. L. Howell, A methodology for systems requirements specification and traceability for large real time complex systems, Technical Report, NAVAL SURFACE WARFARE CENTER SILVER SPRING MD, 1991.

[29] B. Ramesh, M. Jarke, Toward reference models for requirements traceability, IEEE Transactions on Software Engineering 27 (2001) 58–93. doi:10.1109/32.895989.